关于时间序列的金融建模

李程晖

December 9, 2017

1 准备工作

在这门课里，我们对金融数据进行时间序列分析的建模。金融数据本身具有各种性质，比如厚尾性等，同时基本面、消息面等会极大的影响股票的价格，而这些是很难用价格这一个因子去进行刻画的。另外，一只股票的情况受到很多因素的影响，容易出现相当大的波动。不同行业的股票通常具有较大差异的季节向，用一个时间序列去描述很容易出现失真的情况。如果我们对基本面进行参考，同一种基本面因子对于不同行业可能会有着完全不一样的意义。下面我们测试了分行业的股票波动。

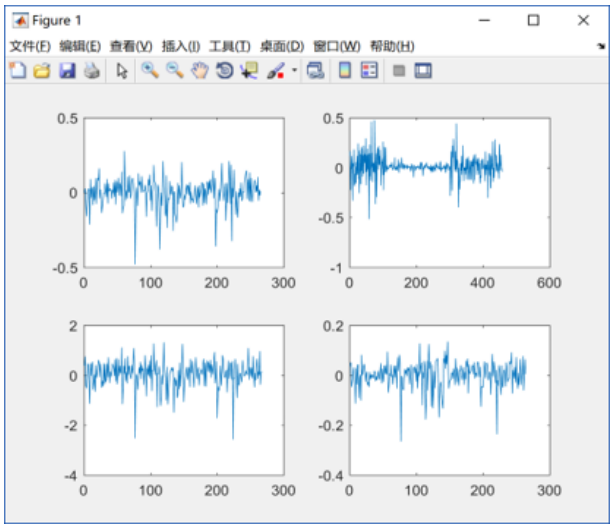

Figure 1: 部分行业测试

Figure 1说的是在同一个周期内不同行业的股票平均价格的波动情况，其中排序分别是行业号为ABCD的行业，其余行业的测试结果略去，具体行业可以参考国民经济行业分类。

我们可以看到实际上股票价格是会有一定的季节向的，而这也是可以理解的。比如农业，农业（序号为A）是以四季为一周期的，当然现在技术上升了可能周期更短，但总而言之会有一个公认的周期，而在这个周期中有一个季节向是比较自然的。而不同行业的周期可能不一样，所以我们不能将整体进行拟合。但是为了简便，我们选择对大盘的数据进行建模。

2 数据

我们找的数据全部来自国泰安的数据网站。

使用数据为2015-12-5到2017-12-5的深圳A股的价格波动情况。先清理数据得到了一共488天开盘的价格波动情况。之后对数据进行建模工作。

先看一下波动情况。Figure 2是放大100倍的价格波动情况。

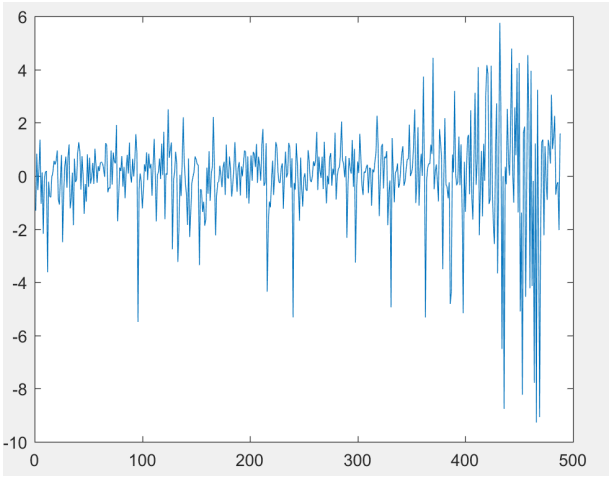

Figure 2: 波动

3 建模

我们先对数据是否为稳定性模型进行检验。

假设模型是$ARMA(p,q)$模型形式：

$$Y_t = \sum_{j=1}^{p} \alpha_j Y_{t-j} + \sum_{j=0}^{q} \beta_j \epsilon_j \tag{1}$$

下面先验证平稳性。

数据的平均值为0.00853。

未差分的数据即Figure 2中的数据。整理成变化曲线的数据未差分的结果为Figure 3。之后将每一个数都减去均值则为中心化的数据。

但是这是股市数据直接减去均值是不合理的，

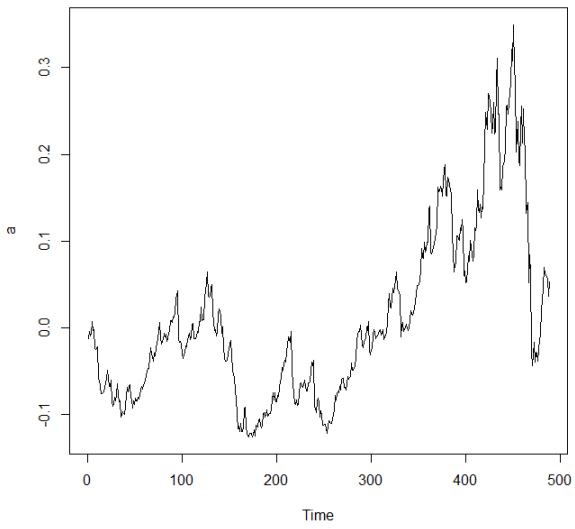

Figure 3: 中心化后

我们应该假设是具有趋势项的，简单处理，假设是二阶函数形式，也即趋势项为$y = at^2 + bt + c$，进行最小二乘法拟合。

$$(a, b, c)^T = (X^T X)^{-1} X^T Y \tag{2}$$

拟合出来的$(a, b, c) = (-2.8e-02, -5.0e-04, 2.0e-06)$，之后拟合出新的曲线。

那么Figure 4 就是新的加入了趋势项的拟合曲线图，图中的二次曲线为趋势项。

下面进行adf检测，得到p-value为0.404，故结果不显著，存在单位根。于是根据结果我们可以认为未差分的序列是不平稳的，之后做一阶差分再次进行adf检验，得到p-value为0.01，结果不显著，存在单位根。故一阶差分结果满足平稳性。

但是一阶差分后数据太小了，不适合做时间序列分析。所以下面依旧是对不平稳的原始数据进行时间序列建模。

根据模型确定具体的$ARMA(p, q)$中模型参数p, q。先画出acf图。

Figure 5即为acf图。

之后画出偏相关图。

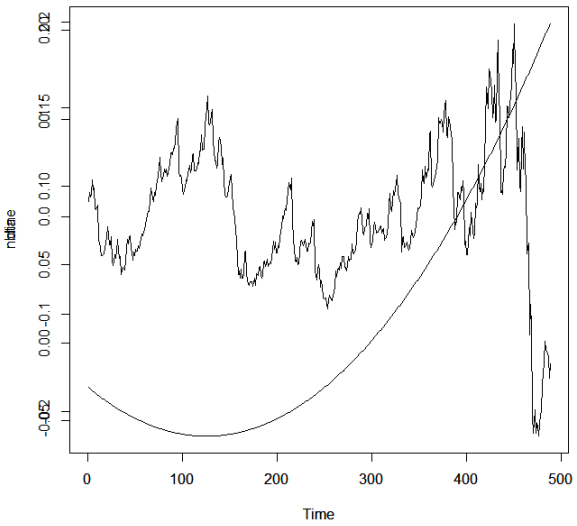

Figure 4: 拟合趋势项

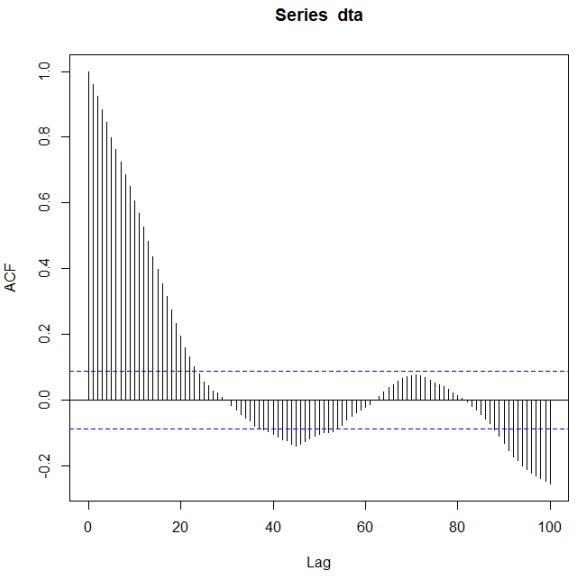

Figure 5: acf

看Figure 5和Figure 6基本可以认为自相关系

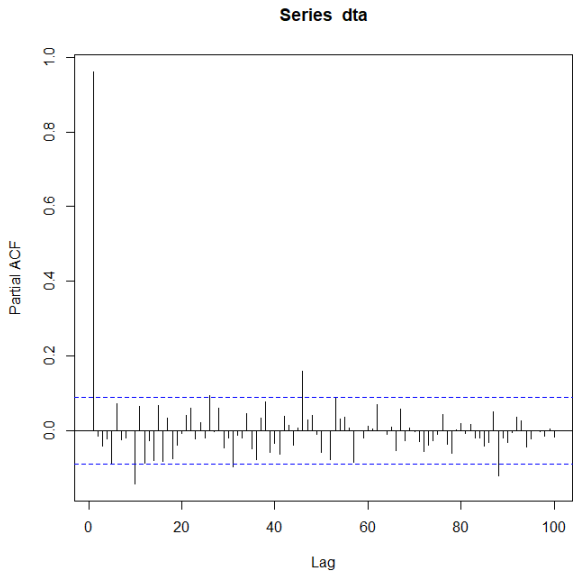

Figure 6: pcf

数是1阶拖尾的，而偏相关系数是截尾的。可以认为数据的形式符合AR(1)模型。同样我们使用R中的auto.arima函数也可以得到这个结论。而图形中的不吻合主要是由于非平稳性所造成的。

那么模型为下述形式，

$$Y_t = b_0 Y_{t-1} + \epsilon_t, \tag{3}$$

之后对假设进行检验$H_0 : Y_t \sim AR(p)$。这个直接看Figure 5即可。

先估计自函数γ_0与γ_1，用公式：

$$\gamma_k = \frac{1}{N}\sum_{j=1}^{N-k} y_j y_{j+k}, k = 0, 1 \tag{4}$$

算出结果：

$$\gamma_0 = 0.00452, \gamma_1 = 0.00434$$

那么由于样本Yule-Walker方程和σ^2的计算方法可以计算出系数与方差：

$$b_0 = \gamma_1/\gamma_0$$

$$\sigma^2 = \gamma_0 - b_0\gamma_1$$

从而可以算出：

$$b_0 = 0.962, \sigma^2 = 0.000173$$

此时样本自回归系数是满足最小相位条件。

$$A(z) = 1 - b_0 z \neq 0, |z| \leq 1 \tag{5}$$

那么我们得到了拟合得到的$AR(p)$模型

$$Y_t = 0.962 Y_{t-1} + \epsilon_t \tag{6}$$

其中$\epsilon_t \sim N(0, 0.000173)$。

下面对模型进行检验工作。

我们对百分误差作平均的结果是60%，之后对残差项的正态性分布进行检验。

假如正确的话，我们拟合得到的残差项应该服从独立正态分布。

需要注意到的一点是根据拟合方式的不同我们可以得到不一样的$AR(p)$模型。假如我们使用内置的R语言的ARIMA的模型获取方法得到模型，我们可以知道模型是这样的：

$$Y_t = 0.9698 Y_{t-1} + \epsilon_t \tag{7}$$

其中$\epsilon_t \sim N(0, 0.0002954)$。

下面我们根据R拟合出来的方程进行检验。首先是正态性检验，我们用QQ图的方式来看正态性。

通过Figure 7两边有一些点不在直线上可以知

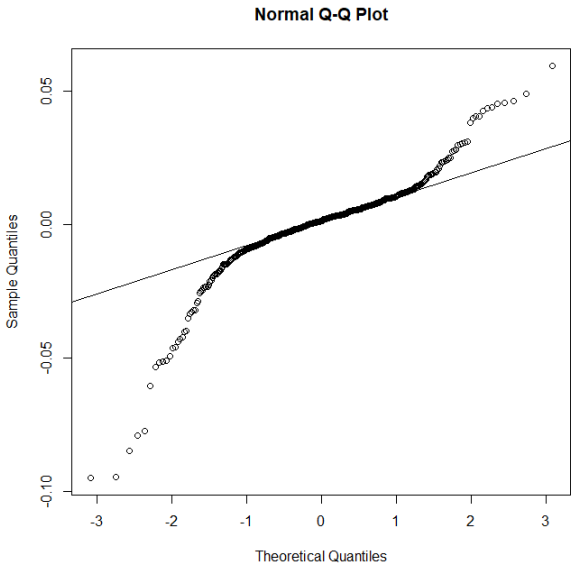

Figure 7: qqnorm

道残差的正态性检验不佳。之后对系数做Ljung-Box显著性检验得到χ^2为0.0047，得到显著性检验的概率为0.9454，从而通过了显著性检验，那么可以认为残差的自相关系数不为零，故ARIMA模型不能满足残差的不相关的假设。

我们可以看一下预测数据的残差图Figure 8。将误差图具体画出来，得到Figure 9。

总而言之，我们可以得到模型为：

$$Y_t - f(t) = Y_{t-1} - f(t-1) + \epsilon_t \tag{8}$$

其中$f(t)$满足：

$$f(t) = 2.8 * 10^{-2} + 5.01 * 10^{-4} * t + 2 * 10^{-6} * t^2$$

整理得到模型:

$$Y_t = Y_{t-1} + 2.8 * 10^{-2} + 4 * 10^{-6} * t + \epsilon_t \quad (9)$$

可以通过时间序列模型得到预测值。Figure 10即预测值，并且同时也给出了置信区间。

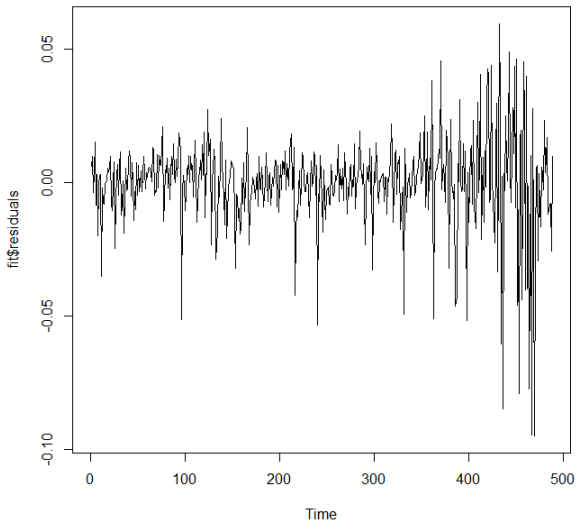

Figure 8: residual1

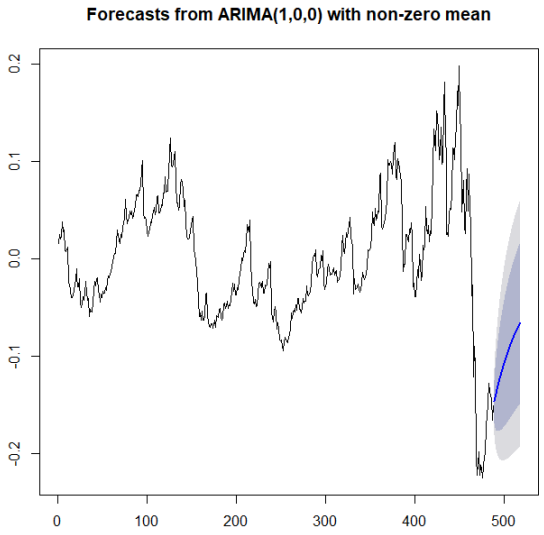

Figure 10: forecast1

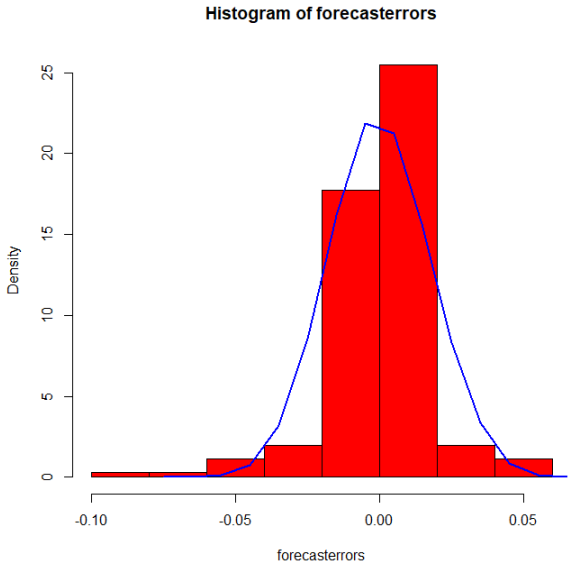

Figure 9: error1

我们来分析原因。一个原因是ß$(x) = 0$ 方程的根为$\frac{1}{b_0}$比较靠近1，所以稳定性不佳。那么此时拟合出来的模型一定是不太稳定的，所以残差不稳定也是可以理解的。而且一开始的数据也有波动。

4 模型改进

假如我们用一阶差分的方法得到模型会稳定一些，拟合程度可能会好一点，我们使用ARIMA（1，0，1）模型进行拟合，QQ残差图如Figure 11，所以残差的正态性检验不佳，Ljung-Box显著性检验的概率为0.0001509，小于0.05，从而知道残差是不相关的，从而模型大致上满足了模型残差的假设。这时候的数据点也即Figure 2。

此时拟合得到的方程为

$$Y_t = -0.5347 Y_{t-1} + \epsilon_t \quad (10)$$

其中$\epsilon_t \sim N(0, 0.0004379)$。

我们可以看到的是，由于ß$(x) = 0$方程的根

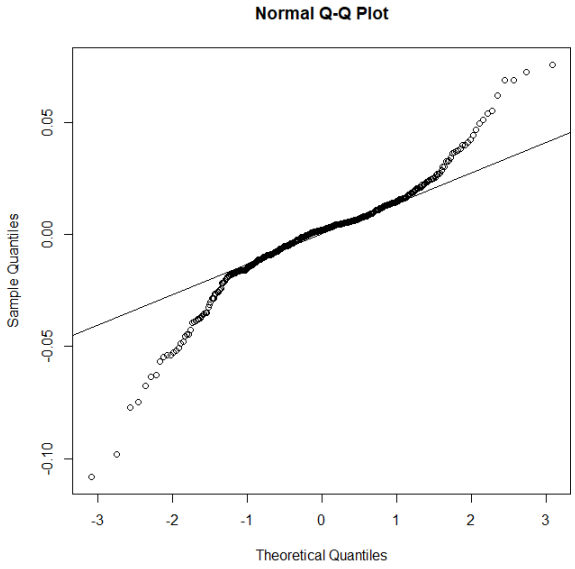

Figure 11: newqqnorm

为$\frac{1}{b_0}$远离1，所以稳定性相比模型9明显加强了。

我们可以看一下预测数据的残差图Figure 12。将误差图具体画出来，得到Figure 13。

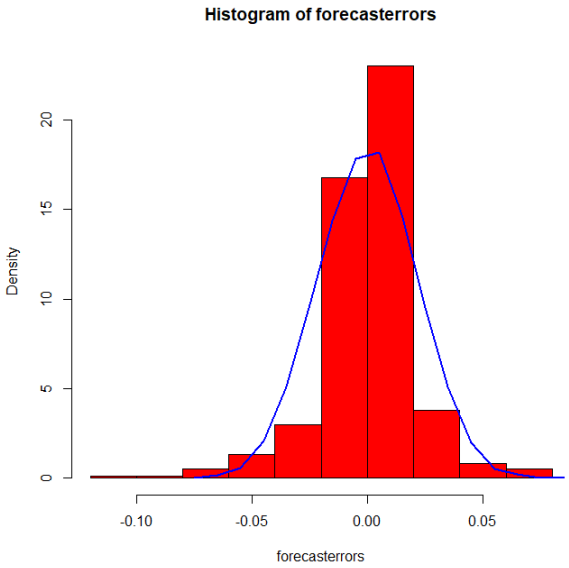

Figure 13: error2

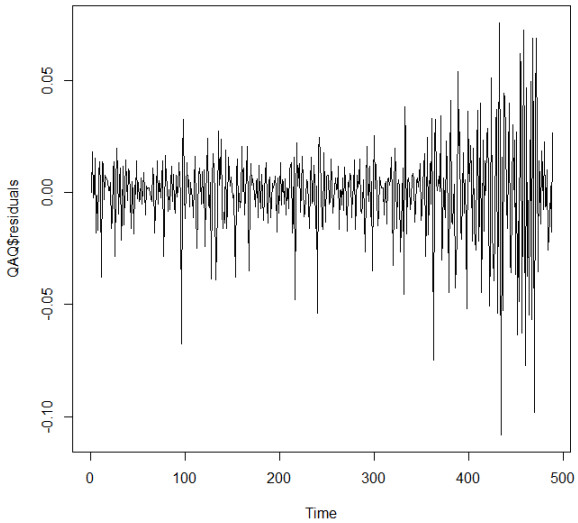

Figure 12: residual2

可以通过时间序列模型得到预测值。Figure 14即预测值，并且同时也给出了置信区间。

从Figure 12与Figure 13可以看到除了部分数据的误差比较大以外，大部分的数据的误差都比较服从正态分布，整个的误差图也大致符合正态分布的样子。

一阶差分后的数据得到的模型为模型10。

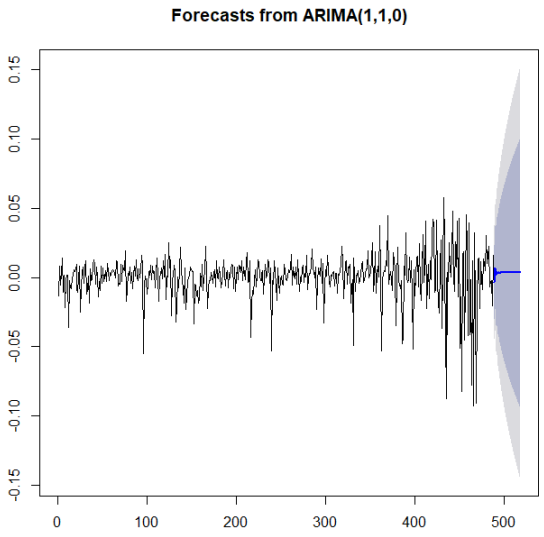

Figure 14: forecast2

5 结论

我们可以通过数据得到一个时间序列模型9，并且可以对模型结果进行测试。若是结果不错，因为σ^2比较小，所以我们可以通过这个模型较准确的进行预测。但是从上述结果可以知道用ARIMA模型拟合深圳市场大盘收益率是欠佳的，一方面拟合得到的ARIMA(0,0,1)模型不稳定，另一方面模型测试的残差不太服从独立性假设和正态性假设。虽然我们可以通过一阶差分得到改进模型10，但是由于数据实际上并不是完全的AR（1）模型，拖尾与截尾的判定整体都有些欠佳，仅仅看ACF和PCF图只能得到模型符合ARIMA(14，14，1)，但这个模型难以加以拟合，并且在实际应用中很有可能过拟合了，并且若是对ARIMA(14,14,1)模型进行拟合的话一定会拟合效果欠佳，所以我们还是拟合AR（1）模型，得到的残差服从独立性假设，但是不太服从正态性假设。从ACF图来看，残差并不是截尾的。

对比模型9与模型10，模型10的稳定性较好，因为单位根离1的距离较模型9的单位根离1的距离较远，而若是离1太近的话，算式：

$$Y_t = \epsilon_t + b_0\epsilon_{t-1} + b_o^2\epsilon_{t-2} + ...$$

难以收敛，而σ^2相对$\frac{1}{b_0}$都比较小。

但是总而言之，数据不太服从ARIMA模型。